Iradxe Bonioxabal Campillo.
02.12.96.

CAPÍTULO V

TABLAS DE CONTINGENCIA Y PRUEBAS DE BONDAD DE UN AJUSTE

5.1. El modelo de referencia

5.2. La hipótesis nula de referencia y la alternativa de asociación

5.3. Estadísticos de contraste

5.4. Determinación de las fuentes de asociación

5.5. Pruebas de homogeneidad de proporciones

5.6. Crítica del análisis de las tablas de contingencia

5.7. Pruebas de bondad de un ajuste

5.8. Prueba chi-cuadrado

5.9. Prueba de Kolmogorov-Smirnov

TEXTO: Apuntes Fotocopiados

BIBLIOGRAFÍA COMPLEMENTARIA:

CANAVOS, G. C.: *Probabilidad y estadística*, Madrid, McGraw-Hill, 1988; capítulo 10, págs. 362-370.

HOEL, P. G.: *Introducción a la estadística matemática*, Barcelona, Ariel, 1976; capítulo 9, págs. 238-254.

TABLAS DE CONTINGENCIA Y PRUEBAS DE BONDAD DE UN AJUSTE: EJEMPLOS

EJEMPLO 1:

El consejo de dirección de una compañía quiere determinar si la opinión Y, de sus accionistas respecto a una propuesta de fusión, es independiente del número de acciones X, que éstos poseen. Una muestra de 500 accionistas proporciona la siguiente tabla:

X \ Y	A favor	En contra	Indecisos	Total
Menos de 200	25	18	21	64
Entre 200 y 1.000	93	62	67	222
Más de 1.000	82	70	62	214
Total	200	150	150	500

Con base en esta información, contrastar a un nivel de significación $\sigma = 0,05$, la hipótesis de que la opinión de los accionistas es independiente del número de acciones que poseen.

EJEMPLO 2: SUPRIMIDO

Durante tres estaciones del año (primavera, verano, otoño) se probaron dos métodos de empaquetado de kiwis, haciéndose un recuento del tanto por ciento de kilos que

llegaron en buen estado a sus destinos. A partir de tres muestras de 200 botes empaquetados con el método A y 150 con el método B respectivamente, se obtuvieron los siguientes porcentajes de kilos defectuosos, según la estación del año:

Porcentaje de kilos	Método A	Método B	Total
Primavera	82	70	152
Verano	25	18	43
Otoño	93	62	155
Total	200	150	350

EJEMPLO 3:

Un proveedor está analizando cuántos pedidos recibe diariamente, y selecciona una muestra aleatoria de 94 días, obteniendo los siguientes resultados:

Nº de días	Nº de pedidos
44	0
22	1
17	2
7	3
3	4
1	5

¿Se puede ajustar a una distribución de Poisson?

(σ = 5%)

EJEMPLO 4:

Para una muestra de 200 personas, se presentan los tiempos necesarios para completar un test.

Tiempo	Nº de personas
20 - 24	15
25 - 29	50
30 - 34	75
35 - 39	40
40 - 44	15
45 - 49	5
	200

Para un σ = 5%, determinar si su distribución es **normal**.

TABLAS DE CONTINGENCIA Y PRUEBAS DE BONDAD DE UN AJUSTE: PROBLEMAS

PROBLEMA 1:

El propietario de una empresa distribuidora de combustibles querría investigar la rapidez con la cual le pagan las facturas en tres zonas diferentes de la ciudad. Los resultados obtenidos en un estudio fueron los siguientes:

	ZONA DE LA CIUDAD		
DÍAS TRANSCURRIDOS HASTA EL PAGO	A	B	C
De 1 a 15	34	42	40
De 16 a 30	48	50	46
Más de 30	18	8	14

Con un nivel de significación del 10%, ¿hay alguna asociación entre la zona y la rapidez con la que pagan las facturas?

PROBLEMA 2:

Se ha efectuado una encuesta para determinar si hay una relación entre el tamaño de la ciudad en la cual reside una persona y la preferencia por un determinado modelo de automóvil. Los resultados de la encuesta se presentan a continuación:

	TIPO DE AUTOMÓVIL				
TIPO DE ÁREA	A	B	C	D	E
Ciudad grande	64	40	26	8	62
Suburbio	53	35	24	6	32
Rural	53	45	30	6	18

Con un nivel de significación del 5%, ¿hay una relación entre el tipo de área de residencia y la preferencia por un modelo de automóvil en concreto?

PROBLEMA 3:

Se efectuó un estudio de mercados a escala nacional, para determinar las preferencias de diversos grupos de edades de varones por los diferentes deportes. A cada varón encuestado se le preguntó cuál era su deporte favorito. Los resultados de la encuesta fueron los siguientes:

	DEPORTE PREFERIDO			
GRUPO DE EDAD	Ciclismo	Fútbol	Baloncesto	Hockey
Menos de 20	26	47	41	36
De 20 a 29	38	84	80	48
De 30 a 39	72	68	38	22
De 40 a 49	96	48	30	26
50 o más	134	44	18	4

Con un nivel de significación del 1%, ¿hay una relación

entre la edad de los hombres y la preferencia por los deportes?

PROBLEMA 4:

Un publicista desea determinar si hay algunas diferencias en tres diferentes medios (TV., radio y prensa) en términos de recordación de un anuncio. Los resultados de un estudio de publicidad fueron los siguientes:

	MEDIO		
CAPACIDAD DE RECORDACIÓN	Prensa	TV.	Radio
Nº de personas que recuerdan el anuncio	25	10	7
Nº de personas que no recuerdan el anuncio	73	93	108

Con un nivel de significación del 10%, determina si hay alguna relación entre los medios y la capacidad de recordar el anuncio.

PROBLEMA 5:

El departamento de estadística de una universidad quiere determinar si hay una relación entre el interés de los estudiantes por la estadística y su capacidad para las matemáticas. El resultado de un estudio realizado entre los alumnos arrojó los siguientes datos:

INTERÉS EN LA ESTADÍSTICA	CAPACIDAD PARA LAS MATEMÁTICAS		
	Baja	Media	Alta
Bajo	60	15	15
Medio	15	45	10
Alto	5	10	25

Con un nivel de significación del 1%, ¿hay relación entre el interés por la estadística y la capacidad para las matemáticas?

PROBLEMA 6:

El responsable del control de calidad en una cadena de producción desea saber si hay una diferencia entre la proporción de piezas defectuosas producidas en diferentes días de la semana laboral. Los resultados de una muestra arrojan estas cifras:

CALIDAD DE PRODUCCIÓN	DÍA DE LA SEMANA				
	Lunes	Martes	Miérc.	Jueves	Viernes
Piezas defectuosas	12	7	7	10	14
Piezas aceptables	86	93	93	90	86

Con un nivel de significación del 5%, ¿hay alguna diferencia en la proporción de piezas defectuosas producidas

en los diversos días de la semana?

PROBLEMA 7:

En una empresa se va a implantar un horario de jornada partida para sustituir al actual de jornada continua. Para conocer la opinión del personal, se lleva a cabo un estudio. Los datos del mismo son:

	ESTADO CIVIL	
OPINIÓN SOBRE LA JORNADA PARTIDA	Soltera/o	Casada/o
La apoyan	63	27
La rechazan	37	23

Con un nivel de significación del 1%, ¿hay pruebas de una diferencia en la actitud hacia el nuevo sistema de jornada en función del estado civil del encuestado?

Resolver el problema utilizando dos métodos distintos.

PROBLEMA 8:

Un grupo de promoción de la mujer está desarrollando una encuesta sobre la presencia femenina en puestos de alta responsabilidad. Los resultados de su estudio son los siguientes:

	SEXO	
PRESENCIA EN	Mujer	Hombre
Puestos de alta responsabilidad	90	125
Puestos de baja responsabilidad	210	175

Con un nivel de significación del 5%, ¿hay diferencia en la proporción de mujeres y hombres que ocupan puestos de alta responsabilidad?

Resolver el ejercicio al menos por dos métodos.

PROBLEMA 9:

Con base en los registros de una tienda de modas, el 50% de los vestidos adquiridos por ésta para la temporada se venderán a precio de menudeo, el 25% a un 20% menos del precio de menudeo, 15% se venderán después de una reducción en su precio del 40% y los restantes con una disminución en su precio del 60%. Para esta temporada, se adquirieron 300 vestidos y su venta fue en la siguiente forma:

Precio de venta	20% de	40% de	60% de
140	90	30	40

¿Existe alguna razón para creer que la disminución en ventas fue diferente en esta temporada con respecto a las

anteriores? Úsese $\alpha = 0,05$. ¿Cuál es el valor de p?

PROBLEMA 10:

Una organización de seguridad vial desea determinar si el número de accidentes fatales se encuentra distribuido de igual forma para el color de los automóviles involucrados en los accidentes. La organización obtuvo una muestra aleatoria de 600 accidentes automovilísticos en los cuales ocurrió, por lo menos, una muerte y anotó el color del automóvil. Se obtuvo la siguiente información:

Rojo	Café	Amarillo	Blanco	Gris	Azul
75	125	70	80	135	115

¿Existe alguna razón para creer que las proporciones de color no son idénticas? Úsese $\alpha = 0,01$.

EJEMPLO 1

DISTRIBUCIÓN CONJUNTA DE FRECUENCIAS OBSERVADAS

MODELO OBSERVADO

cumplen los requisitos necesarios para resolverlo

Y: OPINIÓN QUE POSEEN LOS ACCIONISTAS

CATEGORÍAS: 3

CARACTERÍSTICAS

X \ Y	a favor	en contra	indecisos	Total
200	25 n_{11} (Hay 25 accionistas que están a favor de la fusión y poseen de 200 acc)	18 n_{12}	21 n_{13}	64 n_{14}
entre 200 - 400	93 n_{21}	62 n_{22}	67 n_{23}	222 n_{24}
400	82 n_{31}	70 n_{32}	62 n_{33}	214 n_{34}
Total	200 $n_{\cdot 1}$	150 $n_{\cdot 2}$	150 $n_{\cdot 3}$	500 $n_{\cdot 4}$

CATEGORÍAS: 3

↓

cumplen las características necesarias para su resolución

n_{ij} = $\sum_{i=1}^{n} n_{ij}$: frecuencias conjuntas (de q 'x' tenga l valor c 'y' tenga l valor conjunto l).

frecuencia marginal $n_{i\cdot}$ = $\sum_{i=1}^{m} n_{ij}$

MATRIZ DE PROB. OBSERVADAS

Probabilidades marginales ↓

X \ Y	a favor	en contra	indecisos	Total
[illegible]	P_{11} 0'05	P_{12} 0'036	P_{13} 0'042	$P_{1.}$ 0'128
[illegible]	P_{21} 0'186	P_{22} 0'124	P_{23} 0'134	$P_{2.}$ 0'444
[illegible]	P_{31} 0'164	P_{32} 0'14	P_{33} 0'124	$P_{3.}$ 0'428
Total	$P_{.1}$ 0'4	$P_{.2}$ 0'3	$P_{.3}$ 0'3	1

← Probabilidades marginales

$$P_{ij} = \frac{n_{ij}}{n}$$

probabilidades conjuntas.

$$P_{11} = \frac{25}{500} = 0'05 \qquad P_{12} = \frac{18}{500} = 0'036$$

En [illegible] queremos ver basándonos en nuestra muestra de tamaño "n" contrastar una hipótesis.

→ ¿Cuál sería el modelo teórico de frecuencias observadas si las var. son independientes? Lo obtenemos para contrastarlo con el modelo observado, para poder determinar si las diferencias entre uno y otro son debidas al azar o no.

MATRIZ DE PROB. ESPERADAS

producto de prob. marginales

X \ Y	a favor	en contra	indecisos	Total
200	P'_{11} 0'0512	P'_{12} 0'0384	P'_{13} 0'0384	0'128
500-1000	P'_{21} 0'1776	P'_{22} 0'1332	P'_{23} 0'1332	0'444
1000	P'_{31} 0'1712	P'_{32} 0'1284	P'_{33} 0'1284	0'428
Total	0'4	0'3	0'3	

$$P'_{ij} = P_{i.} \times P_{.j}$$

$P'_{11} = 0'128 \times 0'4 = 0'0512.$

$P'_{12} = 0'128 \times 0'3 = 0'0384$

MODELO TEORICO ESPERADO

X \ Y	a favor	en contra	indecisos	Total
200	E_{11} 25'6	E_{12} 19'2	E_{13} 19'2	64
50 - 100	E_{21} 88'8	E_{22} 66'6	E_{23} 66'6	222
100	E_{31} 85'6	E_{32} 64'2	E_{33} 64'2	214
Total	200	150	150	500

$$E_{ij} = n \times P'_{ij}$$

Pasos para la tabla esperada del modelo teórico (hacer en clase)

$$E_{11} = 500 \times 0'0512 = 25'6$$

$$E_{12} = 500 \times 0'0384 = 19'2$$

Frecuencias esperadas

Resultados esperados para el segundo, pero por ser 2 que falta también se podría.

A continuación comparar el modelo observado con el modelo teórico y ver si las diferencias entre una y otra son significativas o resultado del azar. Para ello construiremos el [illegible]

MODELO DE RESIDUOS

Y	a favor	en contra	indecisos	Total
Sí	n_{11} 25 F_{11} 25'6 R_{11} -0'6	n_{12} 18 F_{12} 19'2 R_{12} -1'2	n_{13} 21 F_{13} 19'2 R_{13} 1'8	0
No	n_{21} 93 F_{21} 88'8 R_{21} 4'2	n_{22} 62 F_{22} 66'6 R_{22} -4'6	n_{23} 67 F_{23} 66'6 R_{23} 0'4	0
NS/NC	n_{31} 82 F_{31} 85'6 R_{31} -3'6	n_{32} 70 F_{32} 64'2 R_{32} 5'8	n_{33} 62 F_{33} 64'2 R_{33} -2'2	0
Total	0	0	0	

$$R_{ij} = n_{ij} - F_{ij}$$

RESIDUOS ESTANDARIZADOS

X \ Y	a favor	en contra	indecisos	Total
[illegible]	- 0'118	- 0'274	0'411	
[illegible]	0'445	- 0'563	0'049	
[illegible]	- 0'389	0'724	- 0'274	
Total				

¡¡ ¿Cómo hallar un residuo estandarizado? !!

$$SR_{ij} = \frac{R_{ij}}{\sqrt{E_{ij}}}$$

E_{ij} frecuencia esperada.

=>

Vamos a comprobar si las diferencias observadas son debidas al azar o no

* ESTADÍSTICO DE CONTRASTE (PEARSON)

la suma de cuadrados de los residuos estandarizados.

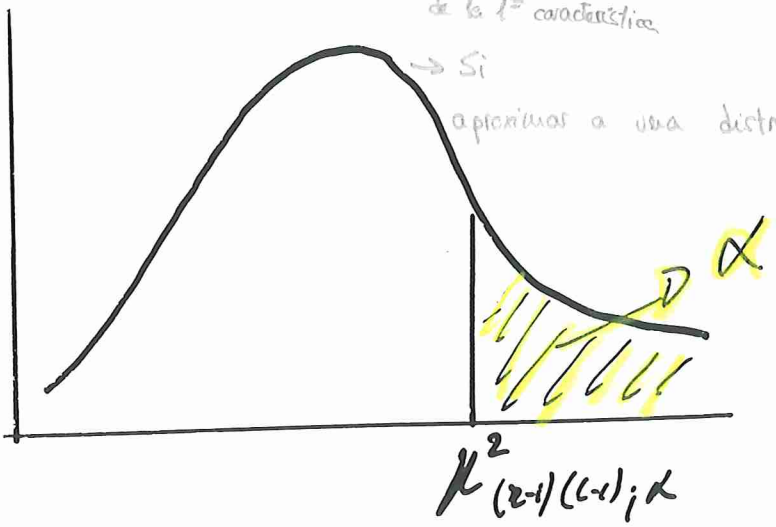
$$EC = \sum_{i=1}^{r} \sum_{j=1}^{c} SR_{ij}^2$$

$EC = 1'525888$ puntuación de

* VALOR CRÍTICO : $\chi^2_{(r-1)(c-1);\alpha}$

r: nº categorías de la 1ª característica

c: nº categorías de la 2 característica



es mayor que podemos aproximar a una distribución normal con "2".

REGLA DE DECISIÓN

→ Si $EC \geq \chi^2_{(r-1)(c-1);\alpha}$ RECHAZAR H_0

Otro método para obtener el estadístico de contraste: el RAZÓN DE VEROSIMILITUD

* RAZÓN DE VEROSIMILITUD

$$EC = 2 \sum_{i=1}^{2} \sum_{j=1}^{c} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right)$$

cociente entre frecuencias observadas y frecuencias teóricas

$$EC = 0'6586$$

* VALOR CRITICO : $\chi^2_{(2-1)(c-1);\alpha}$ ← siempre es este

$$\chi^2_{(3-1)(3-1);0'05} = \chi^2_{4;0'05} = 9'488$$

* REGLA DE DECISIÓN (la misma que la anterior)

Si $EC > \chi^2_{(2-1)(c-1);\alpha}$ RECHAZAMOS H_0

En el ejemplo: NO RECHAZAMOS H_0

INTERPRETACION
Existe evidencia suficiente como para decir que con una prob. de 0'95 las características 'X' e 'Y' son independientes. No existe relación entre la votación y el nº de acciones que posee.

→ ¿Cuándo utilizar uno u otro método?

Cuando las muestras son de tamaño pequeño: mayor fiabilidad el estadístico de contraste de Pearson.

Normal, si con uno de los estadísticos aceptamos una solución, con el otro método tb. (caso: con las muestras pequeñas).

→ En caso de rechazar la hipótesis nula, debemos examinar cuáles son las casillas significativas. Esto se hace según los residuos estandarizados (por defecto →)

TEMA 4

TABLAS DE CONTINGENCIA Y PRUEBAS DE BONDAD DE AJUSTE

1 Tablas de contingencia.

1.1 Concepto: el modelo observado.

1.2 La hipótesis nula de independencia y la alternativa de asociación.

1.3 Estadísticos de contraste.

1.3.1 Para la globalidad de la tabla.

1.3.1.1 El modelo esperado.
1.3.1.2 Los modelos de residuos.
1.3.1.3 Chi cuadrado de Pearson.
1.3.1.4 La razón de verosimilitud.

1.3.2 Para pares de categorías.

1.3.2.1 Las diferencias de proporciones.
1.3.2.2 Los ratios.

1.4 Determinación de las fuentes de asociación.

1.4.1 Análisis de los residuos.

1.4.2 Coeficiente de contingencia.

1.5 Pruebas de homogeneidad de proporciones.

1.6 Crítica del análisis de tablas de contingencia: la paradoja de Simpson.

1.7 Tablas de contingencia y SPSS/PC+

2 Pruebas de bondad de ajuste.

2.1 Concepto y objetivos.

2.2 La prueba de bondad de ajuste Chi cuadrado.

2.3 La prueba de Kolmogorov-Smirnov.

2.4 La prueba de Kolmogorov-Smirnov para dos muestras.

1.- TABLAS DE CONTINGENCIA

1.1 Concepto: el modelo observado

En las aplicaciones de la vida real puede surgir la necesidad de determinar si existe alguna relación entre dos rasgos diferentes en razón de los cuales se ha clasificado una muestra de n observaciones. Ésta es la clásica aplicación de las tablas de contingencia.

Supongamos que nos encontramos con una muestra de n personas a las que hemos clasificado según dos rasgos de interés, por ejemplo, su adicción al tabaco y el posible desarrollo de cáncer pulmonar.

Podemos entonces plantearnos la siguiente cuestión: **¿existe una relación entre el fumar y la predisposición a desarrollar cáncer pulmonar?**

Para cada una de las n personas de la muestra conoceremos su grado de adicción al tabaco (por ejemplo: fumador crónico, fumador moderado y no fumador) y también conoceremos si ha desarrollado o no cáncer pulmonar. Este conocimiento nos permitirá clasificar a las n personas de la muestra según estos dos rasgos.

Cada uno de los rasgos que permiten clasificar la muestra recibe el nombre de **característica**. En nuestro ejemplo, las dos características serán entonces:

* Grado de adicción al tabaco.

* Posible desarrollo de cáncer pulmonar.

Cada una de estas características presenta a su vez una serie de divisiones, que hacen referencia a los distintos valores o estados de naturaleza que puede tomar la característica, y que reciben el nombre de **categorías**. En nuestro ejemplo las categorías son:

* Para la característica "adicción al tabaco" caben tres categorías: "fumador crónico", "fumador moderado" y "no fumador".

* Para la característica "posible desarrollo de cáncer pulmonar" caben dos categorías: "desarrollo de cáncer pulmonar" y "no desarrollo de cancer pulmonar".

Las categorías en que se subdivide cualquier característica deben cumplir dos requisitos:

* Deben ser **exhaustivas** o lo que es lo mismo, deben cubrir todos los posibles estados de naturaleza de la característica a la que pertenecen. La probabilidad de que la característica de interés adopte un estado de naturaleza que no quede recogido en las categorías debe ser nula. En nuestro ejemplo, las categorías de ambas características cumplen este requisito. Respecto a la característica "adicción al tabaco" no cabe otra posibilidad que ser o "fumador crónico" o "fumador moderado" o "no fumador". Respecto al "posible desarrollo de cáncer pulmonar", no cabe otra posibilidad que desarrollarlo o no desarrollarlo.

* Deben ser **mutuamente excluyentes**: es decir, cada observación de la muestra debe pertenecer a una y sólo a una de las categorías de cada característica. Es evidente que en nuestro ejemplo se cumple este requisito: sólo se puede ser o "fumador crónico" o "fumador moderado" o "no fumador" y no se puede ser a la vez "fumador crónico" y "no fumador". Respecto al posible desarrollo de cáncer pulmonar sólo cabe desarrollarlo o no y, lógicamente, no es posible al mismo tiempo desarrollar cáncer pulmonar y no desarrollarlo.

* Cada una de las características deberá tener al menos dos categorías. La explicación intuitiva de este hecho es sencilla: si una característica tiene tan sólo una categoría (esto es, presenta un sólo estado de naturaleza), no supondría ninguna clasificación de las observaciones de la muestra respecto a esa característica, ya que todas ellas se asignarían a la misma categoría (la única), con lo que no tendría sentido tratar de establecer si existe alguna relación entre esa característica (con una sola categoría) y cualquier otra.

* El número de categorías ha de ser finito

Ejemplo: Supongamos que estamos llevando a cabo un estudio en relación a la opinión de las personas respecto a la igualdad de los sexos. Queremos saber si esta opinión se encuentra influenciada por el sexo de los individuos (por ejemplo, sospechamos que las mujeres están en mayor proporción a favor de la igualdad de sexos y queremos comprobar que esto es así).

Para ello clasificaremos la población (que estará formada por el conjunto de todas las personas) en función de las **características de interés** que nos ocupan y que en este caso son "sexo de la persona" y "opinión acerca de la igualdad de sexos".

Así, entre el conjunto de personas de la población habrá algunas a las que respecto a la característica sexo las situaremos en la **categoría** "mujer" y otras a las que situaremos en la **categoría**

"hombre". Vemos fácilmente que estas categorías cumplen los requisitos de ser **exhaustivas y mutuamente excluyentes.** No cabe otra posibilidad que ser hombre o ser mujer (son exhaustivas); no existe ningún estado de naturaleza de la característica sexo que no se encuentre incluído en las dos categorías definidas. Por otra parte, si una persona es hombre, no puede al mismo tiempo ser mujer y viceversa (son mutuamente excluyentes: la pertenencia a una categoría de la característica excluye a la persona de la pertenencia a otra categoría de la misma característica).

Del mismo modo, clasificaremos el conjunto de personas de la población en función de la otra característica de interés, que es su opinión respecto a la igualdad de sexos. De este modo, clasificaremos a algunas dentro de la categoría "a favor de la igualdad de sexos" y a otras (a todas las restantes) en la categoría "en contra de la igualdad de sexos". Estas categorías de la característica "opinión acerca de la igualdad de sexos" son también exhaustivas y mutuamente excluyentes. No cabe otra posibilidad que estar a favor o en contra de la igualdad de sexos (supondremos que toda persona tiene formada una opinión sobre este tema. Si esto no fuera así, deberíamos incluir otra categoría para clasificar en ella a los indecisos). Por otra parte, son mutuamente excluyentes porque si una persona manifiesta estar a favor de la igualdad de sexos no puede al mismo tiempo manifestar estar en contra.

Una vez que hemos dividido la población en función de las características de interés que nos ocupan, debemos comenzar el estudio. Para ello podemos como siempre realizar un censo del conjunto de la población para saber qué parte de ella se clasifica dentro de cada una de las categorías de las dos características definidas. Sin embargo, por motivos ya señalados en temas anteriores, la realización de censos resulta poco menos que imposible. La solución viene dada por la selección de una muestra aleatoria de personas de la población que nos permita estimar qué proporción de personas se encuentran dentro de cada una de las categorías de las características.

Para ello, en nuestro ejemplo acerca de la opinión sobre la igualdad de sexos, tomamos una muestra de 900 personas y las clasificamos respecto a las características de interés, encontrando lo siguiente:

TOTAL MUESTRA: 900 ——> hombres: 497 (55,2%)
——> mujeres: 403 (44,8%)

TOTAL MUESTRA: 900 ——> a favor: 552 (61,3%)
——> en contra: 348 (38,7%)

Vemos que para la característica "sexo" el 55,2% de las personas encuestadas han sido clasificadas dentro de la categoría "hombre" mientras que el 44,8% restante ha sido clasificado dentro de la categoría "mujer". Para la otra característica, "opinión sobre la igualdad de sexos", el 61,3% han sido clasificados dentro de la categoría "a favor", y el 38,7% restante en la categoría "en contra".

Sin embargo, más que las **distribuciones marginales de cada una de las características**, nos interesa la **distribución conjunta** de ambas características:

	MUJERES	HOMBRES	
A FAVOR	254	298	552
EN CONTRA	149	199	348
	403	497	900

Conociendo la distribución conjunta para ambas características, podemos saber que de las 900 personas encuestadas, 254 están simultáneamente clasificadas dentro de la categoría "mujer" relativa a su sexo y dentro de la categoría "a favor" relativa a la característica "opinión respecto a la igualdad de sexos" (es decir, entre las 900 personas encuestadas hay 254 mujeres que están a favor de la igualdad de sexos). Podemos también decir que 199 personas están simultáneamente clasificadas dentro de la categoría "hombre" relativa a su sexo y dentro de la categoría "en contra" relativa a su "opinión respecto a la igualdad de sexos" (es decir, entre las 900 personas encuestadas hay 199 hombres que están en contra de la igualdad de sexos). Similar análisis podemos llevar a cabo para las dos casillas restantes.

Como se puede observar, las distribuciones marginales de cada una de las características de interés aparecen en los márgenes de la tabla. Esto nos permite decir que de los 900 encuestados, 403 son mujeres y 497 hombres y que de los 900 encuestados, 552 están a favor y 348 en contra de la igualdad de sexos.

Cuando se ordenan los datos en una disposición similar a la mostrada, podemos decir que nos encontramos ante una tabla de contingencia o tabla de doble entrada. Esta tabla de contingencia **está formada por las frecuencias que se observaron en la muestra para las dos características y sus correspondientes categorías.**

En general una tabla de contingencia con dos criterios de clasificación (es decir, con dos características de interés) presenta la forma de una matriz de dimensiones rxc, donde "r" es el número de categorías de la característica A y "c" es el número de categorías de la característica B, y cuyos elementos n_{ij} representan el número de observaciones que podemos clasificar simultáneamente en la categoría i-ésima de la característica A y en la categoría j-ésima de la característica B.

La representación genérica de una tabla de contingencia con dos criterios de clasificación es la siguiente:

		CARACTERISTICA B						TOTAL
	categoría	1	2	.....	j	.....	c	
CAR. A	1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	.....	n_{2j}	.....	n_{2c}	$n_{2.}$
	.	.	.	.....	.	.....	.	.
	.	.	.	.....	.	.....	.	.
	i	n_{i1}	n_{i2}	.....	n_{ij}	.....	n_{ic}	$n_{i.}$
	.	.	.	.....	.	.....	.	.
	.	.	.	.....	.	.....	.	.
	r	n_{r1}	n_{r2}	.....	n_{rj}	.....	n_{rc}	$n_{r.}$
	TOTAL	$n_{.1}$	$n_{.2}$	.....	$n_{.j}$	.....	$n_{.c}$	n

Podemos indicar el número total de individuos que presenta la categoría i-ésima de la característica A mediante la siguiente expresión:

$$n_{i.} = \sum_{j=1}^{c} n_{ij} \quad (i = 1, 2, \ldots, r)$$

$$n_{i.} = n_{i1} + n_{i2} + \ldots + n_{ic} \quad (i = 1, 2, \ldots, r)$$

Análogamente, el número total de individuos que presenta la categoría j-ésima de la característica B viene dado por la siguiente expresión:

$$n_{.j} = \sum_{i=1}^{r} n_{ij} \quad (j=1,2,\ldots,c)$$

$$n_{.j} = n_{1j} + n_{2j} + \ldots + n_{rj} \quad (j=1,2,\ldots,c)$$

El número total de individuos en la tabla será entonces:

$$\sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = n$$

Continuando con el ejemplo del estudio sobre la opinión acerca de la igualdad de sexos tendremos que:

$$n_{11}= 254 \; ; \; n_{12}= 298 \; ; \; n_{21}= 149 \; ; \; n_{22}= 199$$

Número de personas que manifiestan estar a favor de la igualdad de sexos:

$$n_{1.} = \sum_{j=1}^{2} n_{1j} = n_{11} + n_{12} = 254 + 298 = 552$$

Número de personas que manifiestan estar en contra de la igualdad de sexos:

$$n_{2.} = \sum_{j=1}^{2} n_{2j} = n_{21} + n_{22} = 149 + 199 = 348$$

Número de mujeres:

$$n_{.1} = \sum_{i=1}^{2} n_{i1} = n_{11} + n_{21} = 254 + 149 = 403$$

Número de hombres:

$$n_{.2} = \sum_{i=1}^{2} n_{i2} = n_{12} + n_{22} = 298 + 199 = 497$$

Número total de personas en la muestra:

$$n = \sum_{i=1}^{2} n_{i.} = \sum_{j=1}^{2} n_{.j} = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} =$$

$$= n_{1.} + n_{2.} = 552 + 348 =$$

$$= n_{.1} + n_{.2} = 403 + 497 =$$

$$= n_{11} + n_{12} + n_{21} + n_{22} = 254 + 298 + 149 + 19$$

$$= 900$$

1.2 La hipótesis nula de independencia y la hipótesis alternativa de asociación

Cuando se analiza una tabla de contingencia se está interesado en contrastar la hipótesis de que las características según las que hemos clasificado la población son independientes.

HIPOTESIS NULA: INDEPENDENCIA

La pertenencia a una categoría de la característica A es independiente de la pertenencia a una categoría de la característica B. O lo que es lo mismo, la probabilidad de que un individuo pertenezca a una determinada categoría de la característica A permanece constante para todas las categorías de la característica B, al tiempo que la probabilidad de que un sujeto pertenezca a una determinada categoría de la característica B permanece constante para todas las categorías de la característica A.

HIPOTESIS ALTERNATIVA: ASOCIACION

Existe alguna asociación entre la pertenencia a una categoría de la característica A y la pertenencia a una categoría de la característica B. La probabilidad de pertenencia a una categoría de la característica A no permanece constante para todas las categorías de la característica B, al tiempo que la probabilidad de pertenencia a una categoría de la característica B no permanece constante para todas las categorías de la característica A.

En nuestro ejemplo del estudio acerca de la opinión sobre la igualdad de sexos, la hipótesis nula supondría que no existe ninguna relación entre el sexo de las personas y la opinión que tienen acerca de la igualdad de sexos. Es decir, la probabilidad de que una persona esté a favor de la igualdad de sexos es la misma para los hombres y para las mujeres. Igualmente equivale a decir que la probabilidad de que una persona esté en contra de la igualdad de sexos es la misma para los hombres y para las mujeres.

Por el contrario, la hipótesis alternativa sugeriría la existencia de una relación entre el sexo de la persona y su opinión acerca de la igualdad de sexos. Según la hipótesis alternativa, la probabilidad de que una persona esté a favor o en contra de la igualdad de sexos no sería la misma para los hombres que para las mujeres.

Para seguir adelante con el análisis, debemos transformar la

matriz de frecuencias observadas en una matriz de probabilidades observadas, mediante la aplicación de las siguientes fórmulas:

$$P_{ij} = \frac{n_{ij}}{n}$$

$$P_{i.} = \sum_{j=1}^{c} P_{ij} = \frac{n_{i.}}{n}$$

$$P_{.j} = \sum_{i=1}^{r} P_{ij} = \frac{n_{.j}}{n}$$

$$\sum_{i=1}^{c} P_{i.} = \sum_{j=1}^{c} P_{.j} = 1$$

Para el ejemplo del estudio acerca de la opinión sobre la igualdad de sexos, la matriz de probabilidades observadas es la siguiente:

	MUJERES	HOMBRES	
A FAVOR	0,282	0,331	0,613
EN CONTRA	0,166	0,221	0,387
	0,448	0,552	1

Nos interesan en primer lugar las probabilidades marginales:

P(a favor) = 0,613
P(en contra) = 0,387
P(mujer) = 0,448
P(hombre) = 0,552

Lo que la hipótesis nula afirma es lo siguiente:

P(a favor/hombre) = P(a favor/mujer) = P(a favor) = 0,613
P(en contra/hombre) = P(en contra/mujer) = P(en contra) = 0,387
P(hombre/en contra) = P(hombre/a favor) = P(hombre) = 0,552
P(mujer/en contra) = P(mujer/a favor) = P(mujer) = 0,448

Podemos todavía expresar la hipótesis nula de otra forma:

$$P(i \cap j) = P(i) \times P(j/i)$$

Si i y j son independientes:

$$P(j/i) = P(j) \Rightarrow P(i \cap j) = P(i) \times P(j)$$

Dado que la hipótesis nula supone independencia entre las características de interés y la hipótesis alternativa supone asociación, podremos expresarlas del siguiente modo:

$$H_o: \text{Independencia: } P_{ij} = P_{i.} \times P_{.j}$$

$$H_a: \text{Asociación: } P_{ij} \neq P_{i.} \times P_{.j}$$

$$i = 1, 2, \ldots, r$$

$$j = 1, 2, \ldots, c$$

Por tanto, bajo el supuesto de independencia que indica la hipótesis nula existirán unas probabilidades esperadas para cada una de las intersecciones entre las categorías de una y otra características. Estas probabilidades reciben el nombre de **probabilidades esperadas** y son las que se espera que existiesen en el caso de independencia entre las características.

En el estudio acerca de la opinión sobre la igualdad de sexos, las probabilidades esperadas serían:

	MUJERES	HOMBRES	
A FAVOR	0,275	0,338	0,613
EN CONTRA	0,173	0,214	0,387
	0,448	0,552	1

Como se puede observar, existen unas diferencias entre las probabilidades observadas (vistas en una tabla anterior) y las probabilidades esperadas (que responden al caso de independencia entre el "sexo" y la "opinión"). La afirmación de la hipótesis nula (la de independencia) es que **las diferencias existentes entre unas y otras probabilidades son debidas al azar.** La aceptación de la hipótesis nula no exige que las diferencias sean nulas, sino que será suficiente con que no excedan un cierto "límite".

El estadístico de contraste propuesto por Pearson es igual a la suma de los cuadrados de todos los residuos estandarizados:

$$EC = \sum_{i=1}^{r}\sum_{j=1}^{c} SR^2_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} \left(\frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}\right)^2 =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \left(\frac{n_{ij} - nP_{ij}}{\sqrt{nP_{ij}}}\right)^2$$

Según Pearson, este estadístico de contraste se distribuye aproximadamente como una Chi-cuadrado con (r-1)(c-1) grados de libertad.

Bajo ciertas condiciones (en concreto, que la frecuencia esperada para cada una de las casillas valga al menos 5) Pearson asume que:

$$\frac{n_{ij} - nP_{ij}}{\sqrt{nP_{ij}}} \Rightarrow Z$$

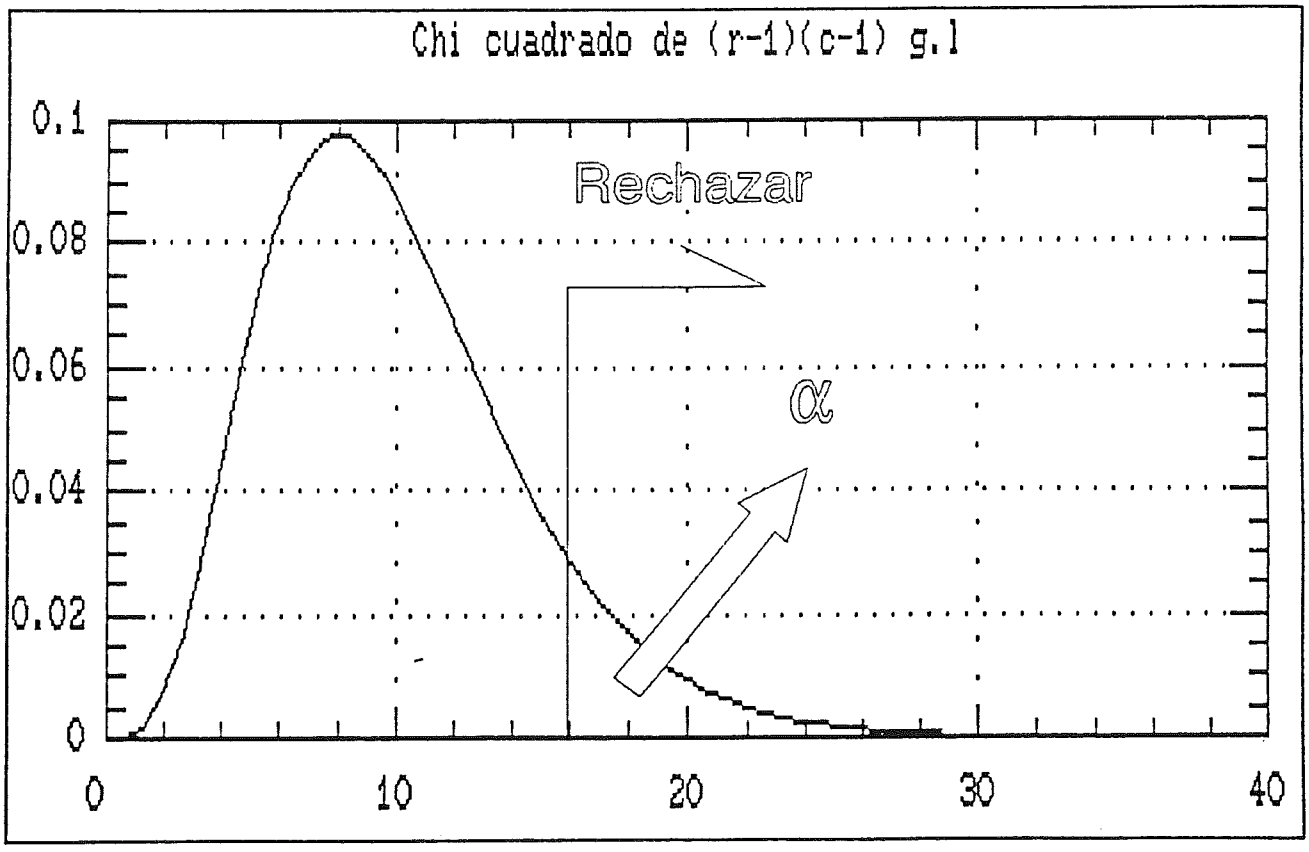
Si esta expresión es N(0,1), la suma de los cuadrados de esta expresión para todas las celdas de la tabla sería una puntuación Chi-cuadrado de rc grados de libertad (recuérdese que la suma de n puntuaciones Z al cuadrado es una puntuación chi-cuadrado de n grados de libertad).

Sin embargo, de los rc sumandos tan sólo (r-1)(c-1) son independientes (concepto de grado de libertad), ya que, respectivamente, r y c sumandos están condicionados por el valor del marginal.

El incumplimiento de la condición establecida por Pearson (todas las frecuencias esperadas valen al menos 5) será tanto más grave **cuanto menor sea el tamaño de la tabla y cuanto menor sea la frecuencia esperada respecto a 5** (la solución a este problema viene dada por el agrupamiento de dos o más categorías de una de las características de interés en una única categoría cuya frecuencia esperada sea al menos de 5).

Si el estadístico de contraste se distribuye según una Chi-cuadrado de (r-1)(c-1) grados de libertad la regla de decisión es clara: se tratará de una prueba unilateral por la derecha, ya que para rechazar la hipótesis nula sólo nos interesan los valores grandes del estadístico de contraste.

El criterio de decisión dirá entonces: cuando el valor del estadístico de contraste sea igual o superior al valor en tablas de la chi-cuadrado de (r-1)(c-1) grados de libertad que deja a su derecha un área α, no aceptaremos la hipótesis nula de independencia; mientras el valor del estadístico de contraste sea menor que el mencionado valor obtenido en las tablas, no podremos rechazar la hipótesis nula de independencia.



Ejemplo:

Se está realizando un estudio para conocer si la preferencia por un determinado refresco está influenciada por la edad de los individuos, o si por el contrario ambas características son independientes. Para ello se ha dividido la población según dos características de interés que son "la preferencia por un determinado refresco" y "la edad".

La característica "preferencia por un determinado refresco" tiene cuatro categorías que son:

* COCA COLA.
* PEPSI COLA.
* KAS.
* TRINARANJUS.

La característica edad tiene tres categorías que son:

* NIÑOS.
* JOVENES.
* ADULTOS.

Se puede ver fácilmente que las categorías de cada una de las características cumplen la condición de ser exhaustivas y mutuamente excluyentes.

Planteamos las siguientes hipótesis:

Hipótesis nula: Existe independencia (o no existe asociación) entre la edad y la preferencia por una cierta marca de refresco. La probabilidad de preferencia de una cierta marca de refresco permanece constante para todas las edades. La probabilidad de pertenencia a un cierto grupo de edad permanece constante sea cual sea el refresco que se prefiera.

Hipótesis alternativa: Existe asociación entre la edad y la preferencia por una cierta marca de refrescos. La probabilidad de preferencia de una cierta marca de refresco no es la misma para todas las edades. La probabilidad de pertenencia a un cierto grupo de edad varía según sea uno u otro el refresco que se prefiera.

Seleccionada una muestra de 200 individuos, hemos encontrado las siguientes frecuencias observadas:

	NIÑOS	JOVENES	ADULTOS	TOTAL
COCA COLA	17	10	13	40
PEPSI COLA	20	10	10	40
KAS	10	20	30	60
TRINARANJUS	15	5	40	60
TOTAL	62	45	93	200

A partir del modelo observado, y basándonos en las frecuencias marginales observadas, construimos el modelo teórico o esperado, que para nuestro caso es el siguiente:

	NIÑOS	JOVENES	ADULTOS	TOTAL
COCA COLA	12,4	9	18,6	40
PEPSI COLA	12,4	9	18,6	40
KAS	18,6	13,5	27,9	60
TRINARANJUS	18,6	13,5	27,9	60
TOTAL	62	45	93	200

Este modelo teórico es el que esperaríamos observar en el caso de que la hipótesis nula de independencia fuese cierta. Comprobemos que bajo este modelo teórico se cumple la hipótesis de independencia:

$$P(NIÑO) = P(NIÑO/COCACOLA) = P(NIÑO/PEPSICOLA) =$$

$$= P(NIÑO/KAS) = P(NIÑO/TRINARANJUS)$$

$$\frac{62}{200} = \frac{12,4}{40} = \frac{12,4}{40} = \frac{18,6}{60} = \frac{18,6}{60} = 0,31$$

$$P(KAS) = P(KAS/NIÑO) = P(KAS/JOVEN) = P(KAS/ADULTO)$$

$$\frac{60}{200} = \frac{18,6}{62} = \frac{13,5}{45} = \frac{27,9}{93} = 0,30$$

Señalemos que se cumple que las frecuencias marginales observadas y esperadas coinciden (como es lógico, ya que las frecuencias esperadas se han calculado en base a las frecuencias marginales observadas).

Digamos que lo ideal para no rechazar la hipótesis nula sería que los modelos observados y esperados coincidiesen plenamente. Sin embargo, vemos que existen diferencias entre ambos, que se concretarán en la obtención del modelo de residuos.

El modelo de residuos en este ejemplo viene dado por:

	NIÑOS	JOVENES	ADULTOS	TOTAL
COCA COLA	4,6	1	-5,6	0
PEPSI COLA	7,6	1	-8,6	0
KAS	-8,6	6,5	2,1	0
TRINARANJUS	-3,6	-8,5	12,1	0
TOTAL	0	0	0	0

En primer lugar, se puede comprobar que los residuos marginales son nulos. Esto se deduce del hecho de que las frecuencias marginales esperadas y las frecuencias marginales observadas coinciden. Por otra parte, conviene interpretar alguno de los residuos. Por ejemplo, de la tabla se deduce que hemos observado 4,6 niños más que prefieren Coca-Cola de los que esperábamos si la hipótesis nula fuese cierta y que hemos observado 8,6 adultos menos que prefieren Pepsi-Cola de los que esperábamos si la hipótesis nula fuese cierta.

Sin embargo, sabemos que estos valores no están ponderados. Son diferencias absolutas que no están relativizadas según el valor de la frecuencia esperada. Para relativizar estos residuos debemos estandarizarlos, lo que se consigue dividiéndolos por la raíz cuadrada de la frecuencia esperada correspondiente:

	NIÑOS	JOVENES	ADULTOS	TOTAL
COCA COLA	1,306	0,333	-1,298	------
PEPSI COLA	2,158	0,333	-1,994	------
KAS	-1,994	1,769	0,398	------
TRINARANJUS	-0,835	-2,313	2,291	------
TOTAL	-----	-----	-----	------

Sumando todos los residuos estandarizados al cuadrado, obtendremos el estadístico de contraste:

$$EC = \sum_{i=1}^{c} \sum_{j=1}^{r} SR^2_{ij} = 30,809$$

Sabemos que este estadístico de contraste se distribuye según una Chi cuadrado de (r-1)(c-1) grados de libertad, que en nuestro caso son 3x2= 6 grados de libertad. Si trabajamos al 95% de nivel de confianza (aunque más frecuente es decir al 5% de nivel de significación) consultaremos la tabla de la Chi cuadrado y buscaremos el valor de 6 grados de libertad que deja a su derecha un área de 0,05. Ese valor es 12,592.

Dado que el valor del estadístico de contraste es mayor que el valor de la Chi-cuadrado de 6 grados de libertad de las tablas concluimos que **rechazamos la hipótesis nula de independencia.** Existe una asociación entre la edad y el refresco preferido. Las diferencias observadas entre el modelo esperado y el modelo observado no se deben al azar y son significativas al 5% (esto quiere decir que existe una probabilidad del 5% de que cuando afirmamos que existe dependencia o asociación nos estemos equivocando).

1.3.1.4 La razón de verosimilitud

Es otro estadístico de contraste para la globalidad de la tabla, que en lugar de basarse en los residuos se basa en el cociente entre las frecuencias observadas y las frecuencias teóricas.

En el caso de la razón de verosimilitud (likelihood ratio), el estadístico de contraste viene dado por:

$$EC = 2\sum_{i=1}^{r}\sum_{j=i}^{c} n_{ij}\ \ln\left(\frac{n_{ij}}{e_{ij}}\right)$$

Aunque no lo demostraremos, este estadístico de contraste sigue también una distribución chi-cuadrado de (r-1)(c-1) grados de libertad, aunque si el tamaño de muestra es reducido lo hace con menos fidelidad que el estadístico de Pearson.

Para nuestro ejemplo de los refrescos, los cocientes entre las frecuencias observadas y las frecuencias esperadas son:

	NIÑOS	JOVENES	ADULTOS	TOTAL
COCA COLA	1,371	1,111	0,699	-----
PEPSI COLA	1,613	1,111	0,537	-----
KAS	0,538	1,481	1,075	-----
TRINARANJUS	0,806	0,370	1,434	-----
TOTAL	-----	-----	-----	-----

Lo ideal para no rechazar la hipótesis nula sería que todos los cocientes fuesen iguales a 1, lo que indicaría que el modelo observado y el modelo esperado coinciden plenamente en todas sus casillas. En ese caso, el estadístico de contraste valdría 0.

Si calculamos el estadístico de contraste según la fórmula indicada, obtenemos que su valor es 32,437.

Como sabemos que este estadístico sigue una distribución Chi-cuadrado de 6 grados de libertad, y que el valor de la chi cuadrado de 6 grados de libertad que deja a su derecha un área de 0,05 es 12,592, nuestra conclusión es que **rechazamos la hipótesis nula de independencia**, con las mismas consecuencias que cuando hemos utilizado el contraste de Pearson.

1.3.2 Estadísticos para pares de categorías

Frente a algunos autores que como Pearson consideran que es posible determinar mediante un estadístico global para toda la tabla la existencia o no de asociación entre las variables de interés, otro grupo de autores consideran que lo único que podemos afirmar es la existencia de relaciones entre categorías de las características. Su opinión es que no se puede resumir en un sólo número la información contenida en una.

Así, estos autores proponen una serie de estadísticos que relacionan pares de casillas y que tratan de determinar las posibles relaciones existentes entre ellas. Vamos a hacer un análisis meramente introductorio y descriptivo de este enfoque.

1.3.2.1 Diferencia de proporciones

Consideremos nuevamente nuestro estudio sobre la actitud de las personas respecto a la igualdad de sexos en relación con su propio sexo. Si recordamos, las frecuencias observadas eran:

	MUJERES	HOMBRES	
A FAVOR	254	298	552
EN CONTRA	149	199	348
	403	497	900

Calculemos la **diferencia de proporciones de sexo sobre la actitud:**

* Proporción de hombres a favor de la igualdad en relación a todos los hombres (298/497) 0,600

* Proporción de mujeres a favor de la igualdad en relación a todas las mujeres (254/403) 0,630

* Diferencia de proporciones: -0,030

Esta diferencia puede interpretarse de dos maneras. Podemos decir que en el colectivo de los hombres hay un 3% menos de personas a favor de la igualdad que en el de las mujeres. También podríamos decir que es 0,03 puntos más probable que una mujer esté a favor de la igualdad de los sexos a que lo esté un hombre.

La diferencia de proporciones tomaría valor 0 cuando no hay asociación y 1 cuando la asociación es perfecta. Cuando no existe asociación, las proporciones serían iguales y por tanto la diferencia entre ambas sería nula. Cuando la asociación es perfecta la proporción de una de las categorías sería la unidad mientras que la proporción de la otra categoría sería nula.

1.3.2.2. Razones o ratios

Existen tres tipos de razones:

* Razones <u>marginales</u>: Se calculan a partir de las frecuencias observadas marginales.

Las razones marginales son el **cociente entre las frecuencias**

observadas marginales de las categorías de una misma característica.

Por ejemplo, en el estudio acerca de la opinión sobre la igualdad de sexo, la razón "hombres-mujeres" viene dada por:

$$\frac{hombres}{mujeres} = \frac{497}{403} = 1,233$$

$$\frac{mujeres}{hombres} = \frac{403}{497} = 0,810$$

Se interpreta diciendo que en la muestra hay 1,233 hombres por cada mujer.

También podemos decir que en la muestra hay 0,810 mujeres por cada hombre.

* Razones condicionales: Se calculan **comparando las frecuencias de una variable entre dos categorías de otra variable.**

Por ejemplo, entre los que están a favor, la razón de "hombres/mujeres" es 1,173 (298/254) mientras que entre los que están en contra, la razón "hombres/mujeres" es 1,335 (199/149).

Si las razones condicionales son significativamente distintas, esto indica que existe asociación entre las categorías de las variables cualitativas objeto del estudio.

* Razón de razones (odds ratio): Hace las funciones de "estadístico de contraste". **Se obtiene dividiendo las razones condicionales.**

En nuestro ejemplo, dividiríamos 1,173 entre 1,335 para obtener un valor de 0,879.

Conclusión: la razón de "hombres/mujeres" entre los que están a favor es menor que esa misma razón entre los que están en contra.

1.4 Determinación de las fuentes de asociación

1.4.1 Análisis de los residuos

Una vez que mediante cualquiera de los estadísticos de contraste analizados hemos llegado a la conclusión de que existe asociación entre las características de interés, podemos estar interesados en conocer cuáles son las categorías de la variable causantes de esa asociación.

Trataremos de localizar aquellas casillas de la tabla que presentan unas **diferencias significativas** respecto a lo que se esperaría haber observado en el caso de independencia.

Sólamente en el caso de que la hipótesis nula haya sido rechazada (sólamente en el caso en que rechacemos la independencia) nos tendremos que preocupar de buscar las casillas que presentan diferencias significativas.

La búsqueda de las fuentes de asociación se lleva a cabo a partir del análisis de los residuos. Sin embargo, ya hemos visto con anterioridad que los residuos "brutos" calculados mediante la simple diferencia entre las frecuencias observadas y las frecuencias esperadas no están adecuadamente relativizados por su frecuencia esperado.

Es por esto que hemos introducido los residuos estandarizados. Aquellas celdas que tengan unos residuos estandarizados elevados supondrán fuentes de asociación.

Los residuos estandarizados siguen una distribución normal y nos interesa tipificarlos para obtener puntuaciones Z. De este modo, aquellos residuos estandarizados y tipificados que superen un cierto valor serán considerados residuos significativos y como tales, fuentes de asociación.

La varianza estimada de los residuos estandarizados viene dada por la siguiente fórmula:

$$V_{ij} = \left(1-\frac{n_{i.}}{n}\right) \cdot \left(1-\frac{n_{.j}}{n}\right)$$

De este modo los residuos estandarizados y tipificados (que reciben el nombre de residuos ajustados de Haberman) se calculan mediante la siguiente expresión:

$$AR_{ij} = \frac{SR_{ij}}{\sqrt{(1-\frac{n_{i.}}{n}) \cdot (1-\frac{n_{.j}}{n})}}$$

Como ejemplo de la utilización e interpretación de los residuos ajustados de Haberman, utilizaremos nuevamente el ejemplo del estudio acerca de la preferencia por un refresco según la edad de los individuos. Para ese ejemplo, los residuos estandarizados de Haberman son los siguientes:

	NIÑOS	JOVENES	ADULTOS	TOTAL
COCA COLA	1,758	0,423	-1,985	-----
PEPSI COLA	2,905	0,423	-3,048	-----
KAS	-2,869	2,402	0,650	-----
TRINARANJUS	-1,201	-3,141	3,743	-----
TOTAL	-----	-----	-----	-----

Dado que los residuos ajustados de Haberman siguen una distribución N(0,1), aquellos valores superiores en valor absoluto a 1,96 serán significativos para un nivel de confianza del 95% y aquellos valores superiores en valor absoluto a 2,575 serán significativos para un nivel de confianza del 99%

Los residuos se interpretan de la siguiente forma:

* AR_{21} = 2,905. Se ha observado un número de niños que prefieren PEPSI que es **significativamente** más elevado de lo que cabría esperar si no hubiera asociación entre la edad y la preferencia por un refresco.

* AR_{31} = -2,869. Se ha observado un número de niños que prefieren KAS que es **significativamente** menor de lo que cabría esperar si no hubiera asociación entre la edad y la

preferencia por un refresco.

* AR_{12} = 0,423. Se ha observado un número de jóvenes que prefieren Coca-Cola superior a lo que cabría esperar si no hubiese asociación entre la edad y la preferencia por un determinado refresco. Sin embargo, la **diferencia no es significativa** y se debe al azar (para un nivel de confianza del 95%)

* AR_{41} = -1,201. Se ha observado un número de niños que prefieren TRINARANJUS inferior a lo que cabría esperar si no hubiese asociación entre la edad y la preferencia por un determinado refresco. Sin embargo, la **diferencia no es significativa** y se debe al azar (para un nivel de confianza del 95%)

No

1.4.2 El coeficiente de contingencia

Es una medida para determinar la intensidad de la asociación a nivel global de la tabla. Se designa con la letra C y viene dado por la siguiente expresión:

$$C = \sqrt{\frac{Valor\ \chi^2\ del\ EC}{Valor\ \chi^2\ del\ EC + n}}$$

C toma siempre un valor entre 0 y 1.

El coeficiente de contingencia no tiene sentido por sí solo y nunca se interpreta en términos absolutos. Sólo tiene sentido cuando lo comparamos con el coeficiente de contingencia de otra tabla de las mismas dimensiones. En este caso, la tabla de contingencia que presente un mayor coeficiente presentará una mayor asociación.

Cuando la tabla de contingencia sea cuadrada, esto es cuando el número de categorías de las dos variables analizadas sea el mismo e igual a k, el máximo valor que alcanza en coeficiente de contingencia será:

$$C_{máximo} = \sqrt{\frac{k-1}{k}}$$

1.5 Las pruebas de homogeneidad de proporciones

Las pruebas de homogeneidad de proporciones son un caso particular de las tablas de contingencia. La resolución de una prueba de homogeneidad de proporciones sigue los mismos pasos que cualquier tabla de contingencia.

La particularidad reside en que bien el número de filas, bien el número de columnas, bien ambos, se reduce a dos: las pruebas de homogeneidad de proporciones son tablas de contingencia de dimensiones 2 x c , r x 2 ó 2 x 2.

El hecho de que al menos una de las características de la tabla de contingencia presente tan sólo dos categorías (sea dicotómica de Boole) nos permite una reformulación de la hipótesis nula: además de expresar la independencia entre las características, podremos hablar de **igualdad de la proporción de éxitos** en c (en el caso de que la tabla tenga 2 filas) o en r (en el caso de que la tabla tenga 2 columnas) poblaciones. La hipótesis alternativa, además de expresarse como la existencia de asociación, se podrá expresar diciendo que **al menos una de las proporciones de éxitos** difiere de las demás.

Ejemplo: Estamos estudiando una fábrica con un horario de trabajo de tres turnos (mañana, tarde y noche), y clasificamos las piezas producidas en cada uno de los turnos según estén bien o mal fabricadas. Hemos tomado una muestra de 100 piezas del turno de mañana, 100 piezas del turno de tarde y 50 piezas del turno de noche y hemos obtenido la siguiente tabla:

	MAÑANA	TARDE	NOCHE	TOTAL
MAL	5	10	10	25
BIEN	95	90	40	225
TOTAL	100	100	50	250

Hipótesis nula: Independencia: No existe asociación entre el turno en el que se ha producido la pieza y la calidad de la pieza. La probabilidad de que una pieza esté bien producida permanece constante para todos los turnos. La probabilidad de que una pieza pertenezca a un determinado turno es la misma tanto si está bien producida como si no lo está.

Ahora bien, si definimos el hecho de que una pieza esté mal producida como éxito y el hecho de que esté bien producida como fracaso y extendemos esto para los tres turnos (esto es, para las tres poblaciones de piezas), tendremos otra forma de plantear la hipótesis nula: la proporción de piezas mal (bien) producidas en el turno de mañana es igual a la proporción de piezas mal (bien) producidas en el turno de tarde y es igual a la proporción de piezas mal (bien) producidas en el turno de noche: igualdad de proporciones poblacionales de éxitos (fracasos).

$$\textit{Una pieza está}$$

$$\textit{mal producida} = \textit{EXITO}$$

$$\textit{bien producida} = \textit{FRACASO}$$

$$H_o : P_{mañana} = P_{tarde} = P_{noche}$$

Hipótesis alternativa: Asociación: existe asociación entre el turno en el que se ha fabricado una pieza y la calidad de la misma. La probabilidad de que una pieza esté bien (mal) fabricada no es la misma para todos los turnos: diferencia en las proporciones poblacionales de éxitos (fracasos), ya que **al menos una de las proporciones poblacionales de éxitos (fracasos) difiere de las demás.**

En nuestro caso, el modelo esperado, el modelo de residuos brutos y el de residuos estandarizados son los siguientes:

	MAÑANA	TARDE	NOCHE	TOTAL
MAL	10 -5 -1,5811	10 0 0	5 5 2,2361	25 0 ---
BIEN	90 5 0,5270	90 0 0	45 -5 -0,7454	225 0 ---
TOTAL	100 0 ----	100 0 ----	50 0 ----	250 0 -----

NO Si calculamos el estadístico de contraste de Person mediante la suma de todos los residuos estandarizados al cuadrado obtenemos:

$$\sum_{i=1}^{r}\sum_{j=1}^{c} SR^2_{ij} = 8,3332$$

$$\chi^2_{2;1-\alpha=0,95} = 5,991$$

En consecuencia, no aceptamos la hipótesis nula. Existe asociación entre el turno en el que se produce la pieza y el hecho de que la pieza esté bien o mal producida. Las diferencias existentes entre el modelo observado y el modelo esperado no se deben al azar y son significativas para un nivel de confianza del 95%

Del mismo modo, al rechazar la hipótesis nula podemos decir que el porcentaje de piezas defectuosas no es el mismo en los tres turnos. Al menos uno de los turnos fabrica con una proporción de piezas defectuosas distinta de los demás.

Sin importar cómo interpretemos el rechazo de la hipótesis nula, debemos señalar que en cinco de cada 100 ocasiones en que afirmemos que existe asociación o que las proporciones de piezas defectuosas no son las mismas para los tres turnos, nos estamos equivocando y en las 95 ocasiones restantes estamos acertando sin saber en cada caso si nos equivocamos o no (probabilidad del error de tipo I).

De todos modos, nos interesará saber cuál es el turno que tiene una proporción de piezas defectuosas distinta a los demás. En otras palabras, nos interesará conocer cuáles son las fuentes causantes de que exista asociación entre el turno en el que se producen las piezas y su validez. Para ello, debemos analizar los residuos ajustados de Haberman, que para el ejemplo que nos ocupa son:

	MAÑANA	TARDE	NOCHE	TOTAL
MAL	-2,152	0	2,635	-----
BIEN	2,152	0	-2,635	-----
TOTAL	-----	-----	-----	-----

De donde se deduce que en el turno de mañana hay un número de piezas defectuosas significativamente menor del que cabría esperar en el caso de que no hubiese asociación entre la validez de la pieza y el turno en el que se ha fabricado, mientras que en el turno de noche hay un número de piezas defectuosas significativamente mayor del que cabría esperar (siempre para un nivel de significación del 5%).

Una observación final: en las tablas de contingencia de dimensiones 2x2, Yates considera que la prueba Chi-cuadrado de Pearson es más exacta si al residuo bruto se le resta media unidad de medida antes de estandarizarlo. Por tanto, el residuo de Yates viene dado por la siguiente expresión

$$R_{ij}\ de\ Yates = R_{ij} - 0.5$$

1.6 Crítica del análisis de tablas de contingencia: Paradoja de Simpson

Las tablas de contingencia llevan implícita la consideración de que los individuos de la población se diferencian tan sólo en aquellas variables que hemos considerado características de análisis y que son las que nos permiten clasificarlos. En los demás aspectos, se asume que los individuos resultan ser homogéneos.

Por ejemplo, en el estudio acerca de la preferencia por un determinado refresco, los individuos de la población se diferencian tan sólo en dos aspectos, su preferencia por un refresco u otro y su edad, siendo en todo lo demás homogéneos (mismo nivel socioeconómico, mismo nivel de estudios,...).

Pero, ¿qué sucede cuando los individuos no son homogéneos? ¿qué sucede cuando nos enfrentamos a un problema real en que los individuos jamás son homogéneos?

Veámoslo en un ejemplo: supóngase que se realiza un experimento para comparar un nuevo tratamiento para una cierta enfermedad con el tratamiento estándar para esa enfermedad. En el experimento se trata a 80 individuos que padecen la enfermedad, de los cuales 40 reciben el tratamiento estándar y los otros 40 reciben el nuevo tratamiento. Después de cierto tiempo se observa cuántos de los individuos de cada grupo han mejorado y cuántos no. Supóngase que

los resultados son los siguientes:

	MEJORAN	NO MEJORAN	TOTAL
NUEVO	20	20	40
ESTANDAR	24	16	40
TOTAL	44	36	80

De acuerdo con esta tabla, 20 de los 40 enfermos que recibieron el nuevo tratamiento mejoraron (el 50%), mientras que 24 de los 40 individuos que recibieron el tratamiento estándar mejoraron (el 60%). A partir de estos datos, **el nuevo tratamiento parece menos efectivo que el tratamiento estándar.**

Pero veamos ahora lo que sucede cuando presentamos separadamente la tabla para los hombres y para las mujeres. Para los hombres la tabla es la siguiente:

	MEJORAN	NO MEJORAN	TOTAL
NUEVO	12	18	30
ESTANDAR	3	7	10
TOTAL	15	25	40

Del estudio de esta tabla se desprende que 12 de los 30 hombres a los que se aplicó el nuevo tratamiento (el 40%) mejoraron, mientras que 3 de los 10 hombres a los que se aplicó el tratamiento estándar mejoraron. De aquí se deduce que **para los hombres el tratamiento nuevo parece ser más efectivo que el tratamiento estándar.**

Para las mujeres la tabla es la siguiente:

	MEJORAN	NO MEJORAN	TOTAL
NUEVO	8	2	10
ESTANDAR	21	9	30
TOTAL	29	11	40

De aquí se deduce que 8 de las 10 mujeres a las que se aplicó el tratamiento nuevo (el 80%) mejoraron, mientras que 21 de 30 mujeres a las que se aplicó el tratamiento estándar (el 70%) mejoraron. Se desprende que **el tratamiento nuevo es más efectivo para las mujeres que el tratamiento estándar.**

Sin embargo, estos resultados son absurdos. Cuando analizamos a hombres y a mujeres separadamente llegamos a la conclusión de que el nuevo tratamiento es, tanto para los hombres como para las mujeres, más efectivo que el tratamiento estándar, mientras que cuando analizamos a hombres y a mujeres conjuntamente resulta que el tratamiento nuevo es menos efectivo que el tratamiento estándar.

Esta situación recibe el nombre de **Paradoja de Simpson** y se debe a la heterogeneidad de los individuos de la población.

Cuando analizamos la tabla de contingencia global, **estamos considerando que los individuos de la población sólo se diferencian por el tratamiento que se les ha aplicado y por su recuperación o no recuperación.** Esto es absolutamente falso. Entre otras diferencias, en esa población hay hombres y hay mujeres y eso supone una gran fuente de heterogeneidad para este análisis, y eso no ha sido tenido en cuenta.

De hecho, comprobemos que la tasa de recuperación de las mujeres con cualquiera de los dos tratamientos es mucho mayor que la tasa de recuperación de los hombres:

* Mujeres con nuevo tratamiento, se recuperan: 80%
* Mujeres con tratamiento estándar, se recuperan: 70%

* Hombres con nuevo tratamiento, se recuperan: 40%
* Hombres con tratamiento estándar, se recuperan: 30%

A pesar de que tanto para hombres como para mujeres la tasa de recuperación con el nuevo tratamiento es mayor que con el tratamiento estándar, podemos conseguir que para un conjunto de individuos formado por hombres y mujeres, la tasa de recuperación para el tratamiento estándar sea mayor que la tasa de recuperación para el tratamiento nuevo.

¿Cómo? Aprovechando que la tasa de recuperación de las mujeres es mayor que la de los hombres (sea cual sea el tratamiento), asignaremos a muchas mujeres y a pocos hombres el tratamiento estándar (con lo que el tratamiento estándar tendrá una gran tasa de recuperación) y asignaremos a pocas mujeres y a muchos hombres el tratamiento nuevo (con lo que la tasa de recuperación para el

nuevo tratamiento será pequeña).

De hecho, si nos fijamos, esto es exactamente lo que se ha hecho en este estudio. De las 80 personas que componen la muestra, el tratamiento estándar se ha aplicado a 10 hombres y a 30 mujeres mientras que el tratamiento nuevo ha sido asignado a 30 hombres y 10 mujeres, consiguiendo así que la tasa de recuperación para el nuevo tratamiento sea menor que la tasa de recuperación para el tratamiento estándar.

Conviene reparar en la posibilidad que cualquier persona tiene de manipular a su conveniencia los datos de una tabla de contingencia. Supóngase que el laboratorio que se encargaba de elaborar el medicamento aplicado en el tratamiento estándar descubre que el tratamiento nuevo, cuyo medicamento no es capaz de producir, resulta más eficaz. Sería muy sencillo para ese laboratorio presentar a todos los hospitales y clínicas del país una tabla de contingencia en la que "demuestre", tomando las muestras que le convenga, que el tratamiento nuevo es menos eficaz que el tratamiento estándar. El laboratorio no falsearía los datos sino que los "utilizaría" del modo que más le conviene.

La paradoja de Simpson demuestra los peligros de realizar inferencias a partir de tablas agregadas. Para garantizar que la paradoja de Simpson no se produzca en nuestro ejemplo, las proporciones relativas de hombres y mujeres que reciben el nuevo tratamiento y el tratamiento estándar deben ser iguales (esto no quiere decir que en la muestra haya el mismo número de hombres que de mujeres). Sin embargo, esto sólo nos garantiza que la paradoja de Simpson no se produce en lo que se refiere al sexo, **pero existen o al menos pueden existir otras muchas fuentes de heterogeneidad.** Supongamos por ejemplo que en la muestra hay pacientes en una fase inicial de la enfermedad (cuya tasa de recuperación podemos suponer más elevada) y pacientes en una fase más avanzada (cuya tasa de recuperación será menor). Nuevamente, la fase en la que se encuentre la enfermedad puede ser una fuente de heterogeneidad.

En definitiva, por muy cuidadosos que seamos en el diseño, la posibilidad de la paradoja de Simpson subyace en cualquier tabla de contingencia.

4.7 Tablas de contingencia y SPSS/PC

El desarrollo de paquetes informáticos ha permitido liberar a las personas de la realización de pesados cálculos. No podemos negar que el análisis de tablas de contingencia, a pesar de no resultar complicado, sí puede suponer la realización de un elevado número de cálculos, lo que conlleva una gran posibilidad de cometer errores.

Entre esos paquetes informáticos se encuentra el SPSS/PC que, entre otras muchas aplicaciones, presenta un subprograma (CROSSTABS) para el análisis de tablas de contingencia. Como ejemplo, veamos la salida del subprograma CROSSTABS para el estudio sobre la preferencia de un refresco según la edad:

```
CROSSTABS /TABLES= REFRESCO BY EDAD
          /CELLS= COUNT EXPECTED RESID SRESID ASRESID
          /STATISTICS= CHISQ CC.
```

		EDAD			
	Count Exp Val Residual Std Res Adj Res	NIÑO	JOVEN	ADULTO	Row Total
REFRESCO		1.0	2.0	3.0	
	1.0	17	10	13	40
COCA COLA		12.4	9.0	18.6	20.0%
		4.6	1.0	-5.6	
		1.3	.3	-1.3	
		1.8	.4	-2.0	
	2.0	20	10	10	40
PEPSI		12.4	9.0	18.6	20.0%
		7.6	1.0	-8.6	
		2.2	.3	-2.0	
		2.9	.4	-3.0	
	3.0	10	20	30	60
KAS		18.6	13.5	27.9	30.0%
		-8.6	6.5	2.1	
		-2.0	1.8	.4	
		-2.9	2.4	.6	
	4.0	15	5	40	60
TRINARANJUS		18.6	13.5	27.9	30.0%
		-3.6	-8.5	12.1	
		-.8	-2.3	2.3	
		-1.2	-3.1	3.7	
	Column Total	62 31.0%	45 22.5%	93 46.5%	200 100.0%

Chi-Square	Value	DF	Significance
Pearson	30.80944	6	.00003
Likelihood Ratio	32.43716	6	.00001
Mantel-Haenszel test for linear association	14.97565	1	.00011
Minimum Expected Frequency	9.000		

2. PRUEBAS DE BONDAD DE AJUSTE

2.1. Concepto y objetivos

Hasta este momento, siempre que hemos realizado estimaciones de parámetros (capítulo 2) o pruebas de hipótesis paramétricas (capítulo 3), hemos supuesto que la muestra de la que disponíamos provenía de una población con una distribución exacta y conocida, aun cuando alguno de los valores de sus parámetros fueran desconocidos. Por ejemplo, suponíamos que la muestra provenía de una población cuya distribución era normal, aunque podíamos desconocer el valor de la media poblacional o el valor de la varianza poblacional.

Las pruebas de bondad de ajuste ofrecen la posibilidad de llevar a cabo inferencias acerca de la naturaleza de la población de la que se ha tomado la muestra. Permiten determinar si una muestra de datos "se ajusta" a una distribución de probabilidad dada (lo que hemos llamado naturaleza) de la población de la que se han extraído. Por ejemplo, mediante las pruebas de bondad de ajuste podremos investigar la posibilidad de que una cierta muestra haya sido tomada de una población cuya distribución es normal.

La hipótesis nula que se plantea en toda prueba de bondad de ajuste consiste en afirmar que la muestra se ha extraído de una población que sigue una cierta distribución. Por contra, la hipótesis alternativa afirma que la población de la que se ha extraído la muestra no sigue esa distribución dada. En base a la información de la muestra, y mediante algún estadístico de contraste, determinaremos si rechazamos o no rechazamos la hipótesis nula.

Veremos en este tema las dos pruebas de bondad de ajuste más conocidas: la prueba Chi-cuadrado de Pearson y la prueba de Kolmogorov-Smirnov.

2.2 La prueba de bondad de ajuste Chi-cuadrada

Es el contraste más antiguo. La idea que subyace bajo el contraste Chi-cuadrado de Pearson es comparar las frecuencias observadas en la muestra con las especificadas por el modelo teórico que se contrasta en la hipótesis nula

El contraste Chi-cuadrado de Pearson es válido para todo tipo de distribuciones, tanto continuas como discretas, si bien en el caso de distribuciones continuas, éstas deberán ser recodificadas en un conjunto limitado de intervalos.

La hipótesis nula supone que unos datos de una cierta variable (que designaremos por X) provienen de un determinado modelo. Evidentemente, la hipótesis alternativa afirma que los datos no provienen de ese modelo.

CONDICIONES A CUMPLIR
$k \geq 5$
$E_i \geq 5$
$n > 25$ ⇒

$$H_o : F(X) = F_o(X)$$

$$F_o(X) \text{ es el modelo teórico}$$

Vamos a comparar las frecuencias observadas en la muestra con las teóricas, las del modelo: veremos si las diferencias son suficiente grandes para rechazar H_0.

Respecto a la forma de la hipótesis nula caben dos posibilidades. En primer lugar, puede suceder que el modelo teórico se encuentre **completamente especificado**, o lo que es lo mismo, conocemos cuál es el tipo de distribución y conocemos todos sus parámetros. Por ejemplo, sabemos que el modelo teórico es normal con media y varianza conocidas o sabemos que el modelo es Poisson con lambda conocida. En segundo lugar, puede suceder que **conozcamos la forma, pero desconozcamos uno o varios de los parámetros**. Por ejemplo, sabemos que el modelo teórico es binomial, pero desconocemos el valor de P. Este desconocimiento afectará a los grados de libertad del estadístico de contraste.

r parámetros desconocidos

χ^2_{k-r-1} (grados de libertad)

En todo caso, el contraste requiere los siguientes pasos:

1º) Agrupar las n observaciones de la muestra en k (finito) clases. Las clases deben ser exhaustivas y mutuamente excluyentes. Es decir, deben elegirse de modo que cubran todo el rango posible de valores de la variable y también de modo que todos los datos

queden clasificados en una cierta clase sin posible ambigüedad. Llamaremos N_i al número de observaciones que quedan clasificadas en la clase i-ésima (modelo de frecuencias observadas)

2º Calcular la probabilidad que el modelo propuesto o teórico de la hipótesis nula asigna a cada una de las k clases que hemos definido. Hay que recordar que la suma de todas las probabilidades deberá ser uno. Designaremos por E_i a la frecuencia esperada según el modelo teórico para la clase i-ésima (modelo de frecuencias esperadas o teóricas)

$$E_i = n \cdot P_i$$

3º Evidentemente, si todas las frecuencias observadas coinciden con las frecuencias esperadas podríamos concluir en principio (ya veremos más adelante que esto es discutible) que la muestra proviene de una población cuya distribución es la indicada en el modelo teórico. Por tanto, tendremos que calcular la discrepancia entre las frecuencias observadas y las frecuencias esperadas según el modelo teórico utilizando la siguiente expresión:

$$EC = \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i}$$

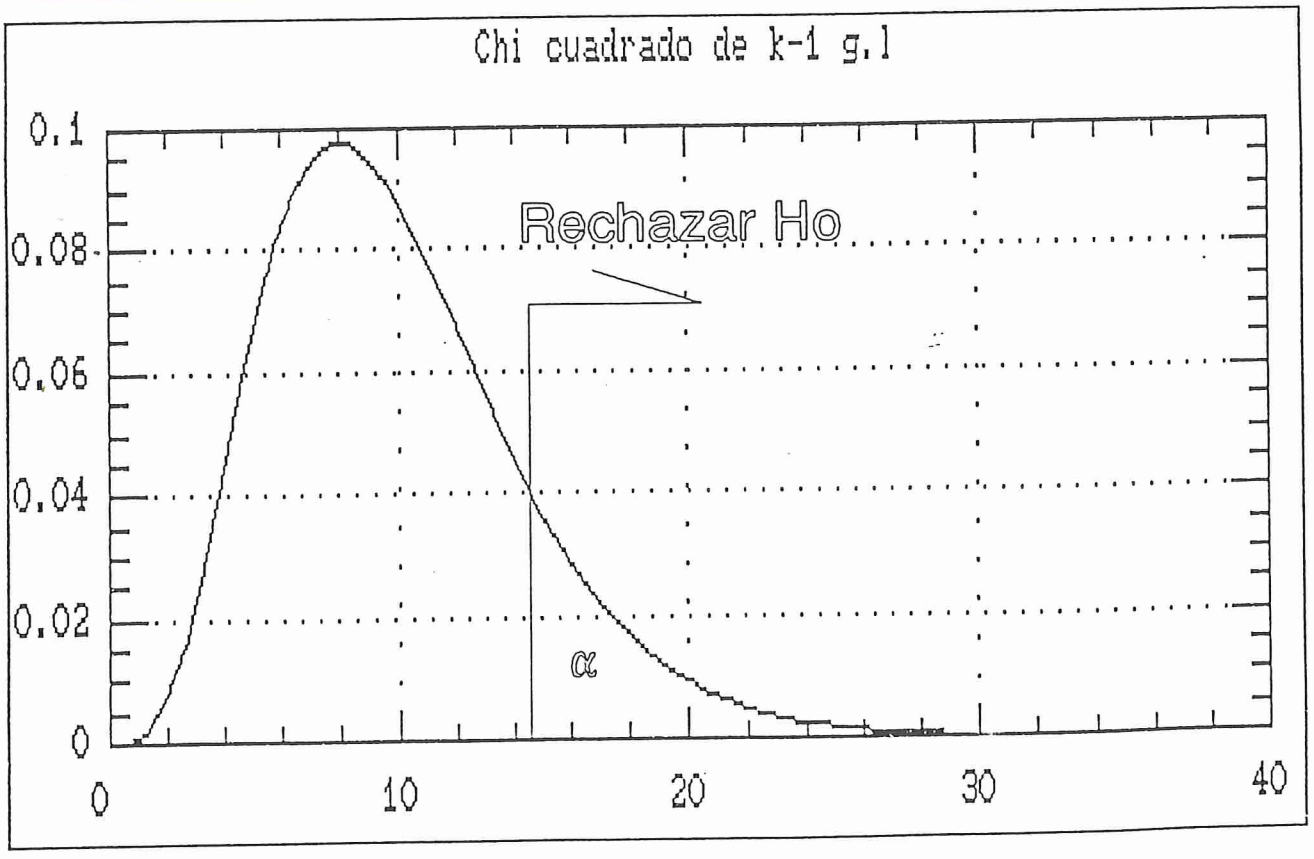
Este estádistico de contraste se distribuye aproximadamente como una Chi-cuadrado. Los grados de libertad dependerán (como ya anunciamos con anterioridad) de si la hipótesis nula estaba planteada completa o parcialmente.

En el caso de hipótesis nula completamente especificada (es decir, conocemos la distribución del modelo teórico y también todos sus parámetros), los grados de libertad del estadístico de contraste serán k-1.

En el caso de hipótesis nula parcialmente especificada (es decir, aquella en la que conocemos la distrbución del modelo teórico pero de la que desconocemos alguno de sus parámetros) los grados de libertad serán k-r-1, donde r expresa el número de parámetros desconocidos que hemos tenido que estimar a partir de la información de la muestra.

En el caso de que la hipótesis nula esté completamente especificada, las probabilidades que el modelo teórico asigna a cada uno de los intervalos no dependen de la muestra que se haya elegido (las probabilidades asignadas a cada intervalo dependerán del modelo y de sus parámetros que son conocidos). Sin embargo, cuando la hipótesis nula no está completamente especificada, las probabilidades que el modelo teórico asigna a cada uno de los intervalos dependen del valor del parámetro o parámetros desconocidos que son estimados a partir de la muestra. Por tanto, en el caso de hipótesis nula parcialmente especificada, las probabilidades teóricas y por tanto las frecuencias esperadas de cada intervalo dependen de la muestra que se haya elegido.

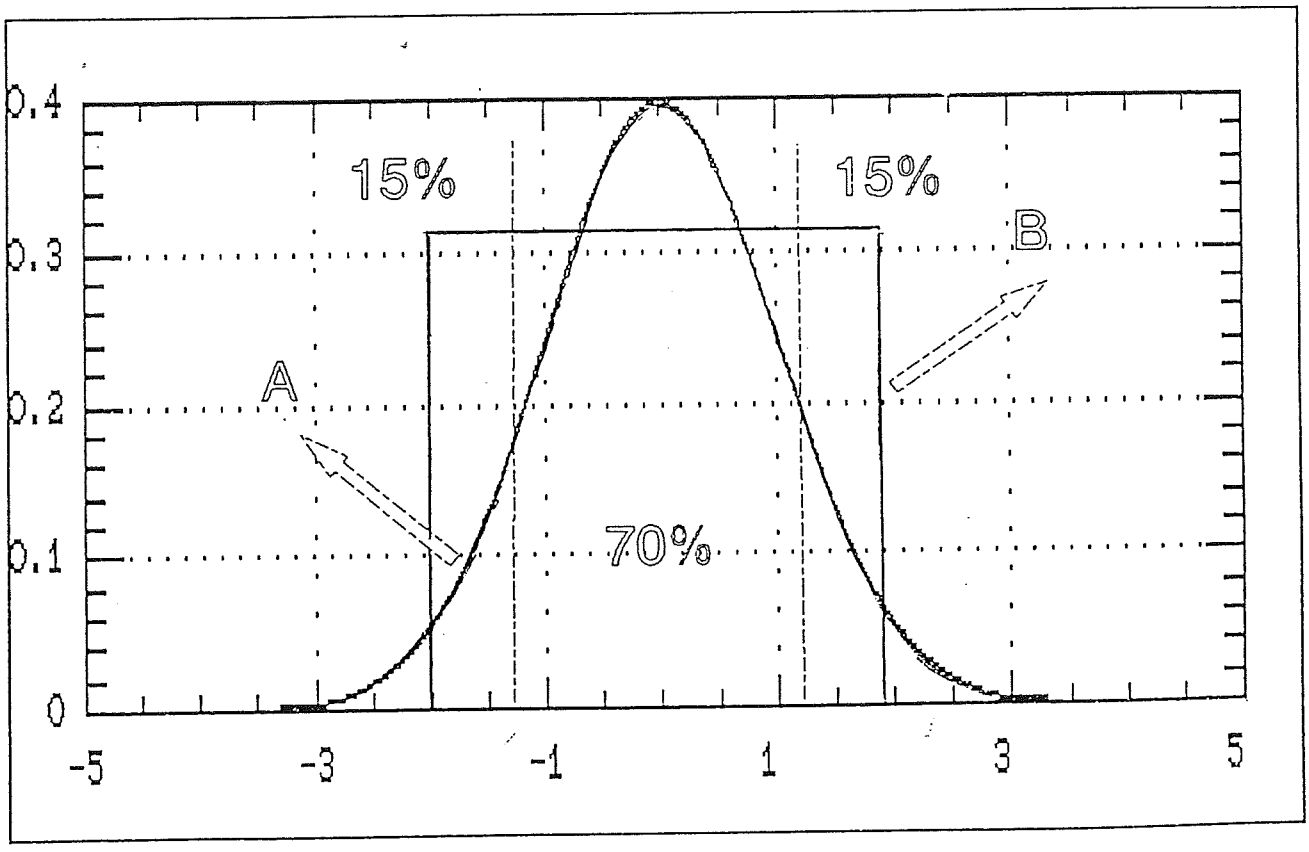
Una vez conocido el estadístico de contraste y su distribución, la regla de decisión está bien clara: cuando el estadístico de contraste supere el valor de la Chi cuadrado de k-1 o k-r-1 grados de libertad (según la hipótesis nula esté o no esté completamente especificada) que deja a su derecha un área α, rechazaremos la hipótesis nula y diremos que la muestra no ha sido elegida de una población que siga el modelo teórico especificado. Por el contrario, cuando el estadístico de contraste no supere el valor mencionado, no rechazaremos la hipótesis nula y diremos que no tenemos pruebas para rechazar que la muestra haya sido seleccionada de una población cuya distribución coincide con la especificada.



Como vimos con anterioridad, el empleo del estadístico de contraste de Pearson exige que las frecuencias esperadas para todas y cada una de las clases sean iguales o superiores a cinco.

La exigencia de que las frecuencias esperadas sean al menos iguales a cinco es uno de las limitaciones más importantes del estadístico de contraste de Pearson.

Otro problema que se plantea es que, dada su forma, el test no contrasta un modelo concreto con la muestra sino que contrasta todos los modelos que atribuyen probabilidades iguales a los intervalos que se han construido. Veamoslo en un gráfico:



Los modelos A y B son indistinguibles en un contraste Chi-cuadrado con tres clases ya que asignan idénticas probabilidades a los intervalos definidos. Por tanto, una muestra que se ajuste adecuadamente al modelo B, se ajustaría (con el mismo valor del estadístico de contraste) al modelo A.

Para evitar esta eventualidad conviene que el número de intervalos definidos sea suficientemente grande. Cuando el número de intervalos es grande, los modelos que asignan idénticas

probabilidades a los intervalos definidos son muy parecidos. En la práctica, es recomendable que el número de intervalos sea al menos de 5.

Si por un lado estamos exigiendo que la frecuencia esperada para cada uno de los intervalos sea al menos de cinco y por otra parte estamos recomendando que el número de clases sea también como poco de 5, el estadístico de contraste Chi-cuadrado de Pearson exige tamaños de muestra superiores a 25 ó 30 observaciones.

Ejemplo: Durante la segunda guerra mundial se dividió el mapa de Londres en cuadrículas de 0,25 Km² y se contó el número de bombas caídas en cada cuadrícula durante un bombardeo alemán. Los resultados fueron:

Impactos:	0	1	2	3	4	5
frecuencia:	229	211	93	35	7	1

Contrastar la hipótesis de que los datos siguen una distribución de Poisson.

Hipótesis nula: La variable X: "Número de impactos en 1 cuadrícula durante el bombardeo" sigue una distribución de Poisson. (la hipótesis nula está incompleta, ya que desconocemos su parámetro lambda que deberemos estimar).

Hipótesis alternativa: La variable X: "Número de impactos en 1 cuadrícula durante el bombardeo" no sigue una distribución de Poisson.

En primer lugar, dado que desconocemos el parámetro lambda, tendremos que estimarlo con los datos de la muestra:

$$\lambda = \frac{\sum X_i f_i}{\sum f_i} = \frac{535}{576} = 0,929$$

Las frecuencias esperadas especificadas por el modelo teórico son:

$$E_0 = 227,5 \quad E_1 = 211$$

$$E_2 = 98 \quad E_3 = 30$$

$$E_4 = 7 \quad E_5 = 1,5$$

Como la frecuencia esperada para el intervalo 6 (5 impactos en la cuadrícula) es menor de 5, tendremos que convertir el intervalo 5 y el 6 en uno solo cuya frecuencia observada y frecuencia esperadas serán:

$$f_{5,6}=8 \qquad E_{5,6}=8,5$$

Por tanto, dado que conocemos las frecuencias observadas y las frecuencias esperadas, podemos calcular el valor del estadístico de contraste:

$$EC=\frac{(229-227,5)^2}{227,5}+\ldots+\frac{(8-8,5)^2}{8,5}=1,13$$

Los grados de libertad del estadístico de contraste serán 5-1-1 (ya que hemos estimado un parámetro). El valor de la Chi cuadrado de 3 grados de libertad que deja a su derecha un área de 0,05 es 7,81.

En consecuencia no tenemos pruebas para rechazar la hipótesis nula. No tenemos pruebas para afirmar que la muestra no provenga de una distribución de Poisson con lambda 0,929. Las diferencias existentes entre las frecuencias observadas y las frecuencias esperadas se deben al azar y no son significativas para un nivel de confianza del 95%

2.3 La prueba de Kolmogorov-Smirnov

La prueba de bondad de ajuste por el método Chi-cuadrado exigía que las frecuencias esperadas en cada uno de los intervalos ó categorías definidos fuese mayor que cinco, al tiempo que recomendaba que los intervalos o categorías definidos por el modelo observado fuesen suficientemente numerosos (un mínimo recomendado de 5). Esta doble exigencia conducía a que el tamaño de muestra debía ser bastante elevado (alrededor de 25 ó 30).

Pero, ¿qué sucede en el caso de que tengamos una variable continua y una muestra de tamaño reducido? Para solucionar este problema existe el **contraste de bondad de ajuste de Kolmogorov-Smirnov.**

La prueba de Kolmogorov-Smirnov es la más apropiada en el caso de que nos encontremos con una variable continua. No necesita que los datos se encuentren agrupados en intervalos y no requiere grandes tamaños de muestra.

La prueba de Kolmogorov-Smirnov se basa en una comparación entre la función de distribución que se observa en la muestra ordenada y la función de distribución propuesta en la hipótesis nula. Si la comparación revela diferencias sustanciales, la hipótesis nula se rechaza.

El procedimiento para llevar a cabo el contraste es:

1º) Ordenar los valores muestrales de modo que:

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \ldots \leq X_{(n)}$$

2º) Calcular la función de distribución empírica de la muestra, a la que desginaremos por Sn(X), y que tendrá la siguiente forma:

$$0 \; si \; X < X_{(1)}$$

$$\frac{r}{n} \; si \; X_{(r)} \leq X < X_{(r+1)}$$

$$1 \; si \; X \geq X_{(n)}$$

En otras palabras, para cualquier valor ordenado X de la muestra aleatoria, Sn(X) es la proporción del número de valores en la muestra que son iguales o menores a X.

3º) Dado que conocemos las funciones de distribución empírica y teórica, debemos calcular la discrepancia entre ellas. Para ello seguiremos el siguiente camino: para cada una de las observaciones ordenadas de la muestra, compararemos el valor de la función de distribución teórica con el valor de la función de distribución empírica para la observación actual y para la anterior, seleccionando la máxima de las dos diferencias.

Una vez llevado a cabo este proceso para todas las observaciones de la muestra, habremos obtenido un conjunto de n diferencias, de las cuales seleccionaremos la mayor. Es decir, será la máxima de las diferencias máximas la que nos interese, la que constituirá el estadístico de contraste de la prueba de bondad de ajuste de Kolmogorov-Smirnov. Esquemáticamente:

Para cada observación:

$$máx[\,|S_n(X_{r-1}) - F_0(X_r)|\,,\,|S_n(X_r) - F_0(X_r)|\,]$$

Tendremos n diferencias.

Seleccionamos la máxima.

Las tablas para este estadístico de contraste ofrecen valores críticos para un tamaño de muestra determinado y para un cierto nivel de confianza. Si este estadístico de contraste supera el valor crítico que nos ofrecen las tablas, no aceptaremos la hipótesis nula. En caso contrario, no la rechazaremos.

Ejemplo: A continuación se proporcionan los valores de una muestra aleatoria del número de respuestas correctas para un determinado test que se aplicó a todos los estudiantes que ingresaron en la universidad: 957, 852, 1048, 910, 981, 933, 963, 1007, 1035, 1010, 1015, 875, 1018, 1063, 1023, 998. En años anteriores, el número de respuestas correctas estaba representado, en forma adecuada por una distribución normal con media 985 y desviación estándar 50. Con base en esta muestra, ¿existe alguna razón para creer que ha ocurrido un cambio en la distribución de respuestas correctas para el test? Empléese un α=0,05.

Sea X: Número de respuestas correctas obtenidas por 1 alumno en el test de acceso a la universidad.

Hipótesis nula: La muestra aleatoria ha sido extraída de una población que se distribuye normalmente con media 985 y desviación estándar 50 (modelo completo).

$$H_o : F(X) \Rightarrow N\ (\mu_x=985\ ;\ \sigma_x=50)$$

Hipótesis alternativa: La muestra aleatoria no ha sido seleccionada de una población que se distribuye normalmente con media 985 y desviación estándar 50:

$$H_o : F(X) \neq N\ (\mu_x=985\ ;\ \sigma_x=50)$$

Veamos cuales son las diferencias para cada una de las observaciones:

	Valores	Sn(x)	Fo(X)	Sn(X)-Fo(x)	Sn(X_1)-Fo(x)	Máx.dif.
1	852	0.0625	0.0039	0.0586	0.0039	0.0586
2	875	0.1250	0.0139	0.1111	0.0486	0.1111
3	910	0.1875	0.0668	0.1207	0.0582	0.1207
4	933	0.2500	0.1492	0.1008	0.0383	0.1008
5	957	0.3125	0.2877	0.0248	0.0377	0.0377
6	963	0.3750	0.3300	0.0450	0.0175	0.0450
7	981	0.4375	0.4681	0.0306	0.0931	0.0931
8	998	0.5000	0.6026	0.1026	0.1651	0.1651
9	1007	0.5625	0.6700	0.1075	0.1700	0.1700
10	1010	0.6250	0.6915	0.0665	0.1290	0.1290
11	1015	0.6875	0.7257	0.0382	0.1007	0.1007
12	1018	0.7500	0.7454	0.0046	0.0579	0.0579
13	1023	0.8125	0.7764	0.0361	0.0264	0.0361
14	1035	0.8750	0.8413	0.0337	0.0288	0.0337
15	1048	0.9375	0.8962	0.0413	0.0212	0.0413
16	1063	1.0000	0.9406	0.0594	0.0031	0.0594

Vemos de este modo que la máxima de las diferencias máximas es 0,17. Este es nuestro estadístico de contraste. Si buscamos en las tablas el valor de la diferencia máxima para un $\alpha=0,05$ y un tamaño de muestra de 16 vemos que es 0,328. Dado que el valor obtenido a partir de los datos de la muestra es inferior al valor crítico de las tablas, no rechazamos la hipótesis nula. No tenemos pruebas para afirmar que la muestra no haya sido seleccionada de una población que se distribuya normalmente con media 985 y desviación estándar 50. No tenemos pruebas para suponer que la distribución del número de respuestas que acierta una persona haya variado respecto a los años anteriores. Las diferencias observadas entre la función de distribución empírica (u obtenida en la muestra) y la función de distribución teórica se deben al azar y no son significativas para un nivel del confianza del 95%

3.4. La prueba de Kolmogorov-Smirnov para dos muestras

La prueba de Kolmogorov-Smirnov para dos muestras trata aquellos problemas en los que se selecciona una muestra aleatoria de m observaciones de una población cuya distribución es desconocida y se selecciona otra muestra aleatoria **independiente** de n observaciones de otra población cuya distribución es también desconocida.

Se desea contrastar el hecho de que ambas poblaciones se distribuyen según el mismo modelo, frente a la hipótesis alternativa de que se distribuyen siguiendo modelos distintos.

$$H_o: F(X) = G(X)$$

$$H_a: F(X) \neq G(X)$$

Para llevar a cabo este contraste los pasos a seguir son los siguientes:

1º Calcular las funciones de distribución empíricas u observadas para cada una de las muestras. Estas funciones de distribución empíricas las denotaremos por Fm(X) y Gn(X). Expresarán para cada una de las observaciones de cada una de las muestras la proporción de observaciones menores o iguales a ella.

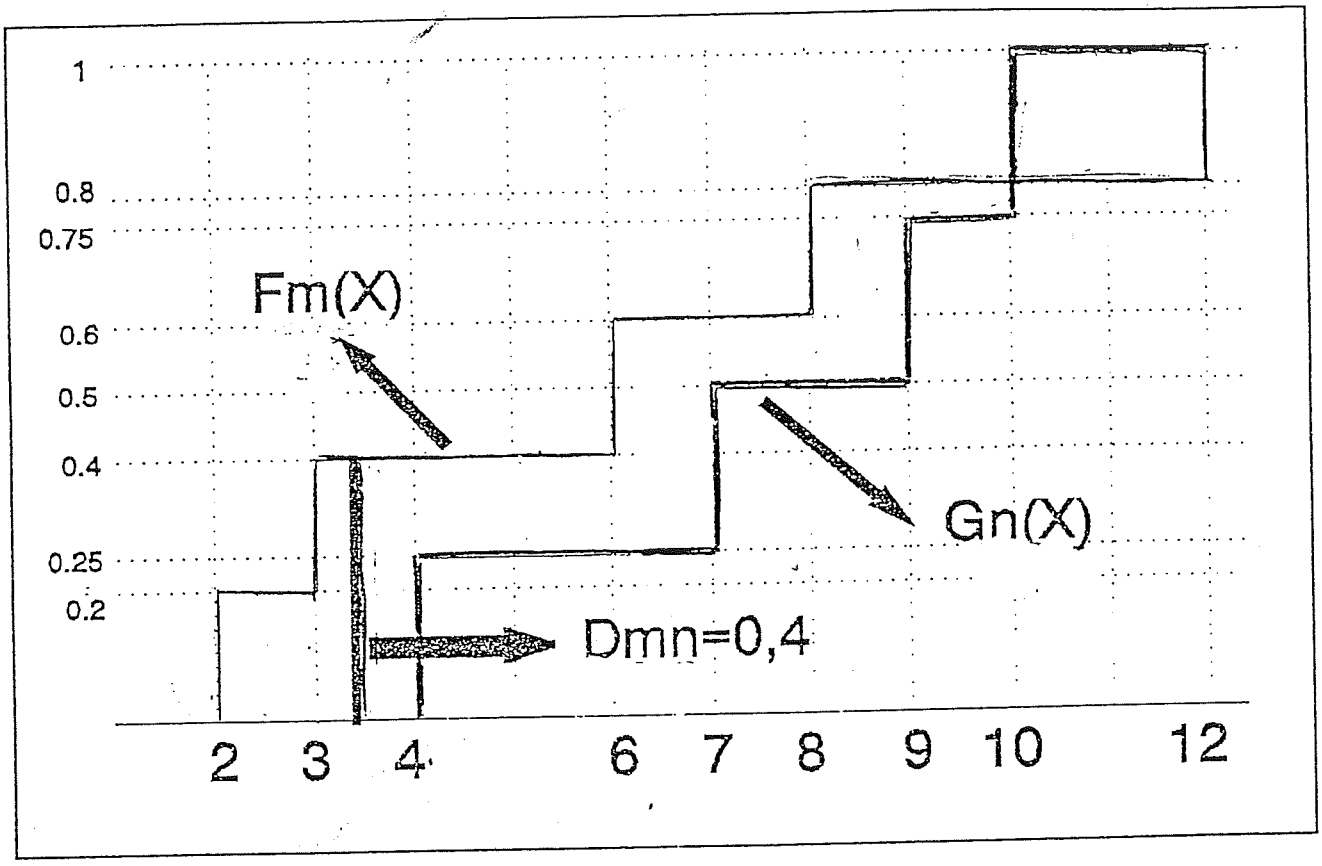
2º. Graficar las funciones de distribución empíricas para cada una de las muestras.

3º. Determinar la máxima diferencia entre las funciones de distribución empíricas para cada una de las muestras. Esta diferencia máxima la denotaremos por Dmn.

La diferencia máxima constituirá el estadístico de contraste de la prueba Kolmogorov-Smirnov para dos muestras independientes. Cuando la diferencia máxima (el estadístico de contraste) sea muy grande, no aceptaremos la hipótesis nula de que ambas muestras proceden de poblaciones con idéntica distribución, mientras que en el caso de que la diferencia sea pequeña no rechazaremos la hipótesis nula.

La regla de decisión es muy sencilla. Se compara la máxima diferencia que hemos obtenido entre las funciones de distribución empíricas de las muestras con la diferencia crítica que aparece en tablas para los valores m y n de tamaño de muestra y para un nivel de confianza dado. De ese modo decidiremos si rechazamos o no rechazamos la hipótesis nula.

Para tamaños de muestra grande, se puede aproximar Dmn mediante la siguiente expresión:



Vemos que el valor del estadístico de contraste es 0,4. Lo que debemos hacer a continuación es comparar la diferencia máxima con el valor que para unos tamaños de muestra 5 y 4, y para el nivel de confianza deseado (en este caso el 95%), nos ofrecen las tablas. Debido a la carencia de tablas para este contraste, supondremos que los tamaños de muestra son suficientemente grandes como para aplicar la aproximación:

$$D_{mn\ máx} = \frac{1,36}{\sqrt{\frac{20}{9}}} = 0,9123$$

Si comparamos el valor del estadístico de contraste que hemos obtenido a partir de las muestras con la máxima discrepancia permitida, decidiremos que no podemos rechazar la hipótesis nula.

No tenemos pruebas para negar que ambas muestras procedan de la misma población. Las diferencias observadas entre las funciones de distribución se deben al azar y no son significativas al nivel del 5%